DOCUMENT RESUME

ED 076 658 TM 002 665

AUTHOR	Besel, Ronald
TITLE	Using Group Performance to Interpret Individual Responses to Criterion-Referenced Tests.
PUB DATE	Feb 73
NOTE	10p.; Paper presented at Annual Meeting of American Educational Research Association (New Orleans, Louisiana, February 25-March 1, 1973)
EDRS PRICE	MF-$0.65 HC-$3.29
DESCRIPTORS	*Criterion Referenced Tests; *Mathematical Models; *Measurement Instruments; Speeches; Statistical Analysis; *Test Interpretation; *Test Results
IDENTIFIERS	*Mastery Learning Test Model

ABSTRACT

The contention that interpretation of a student's performance on a criterion referenced test should be independent of the performance of his classmates is challenged. The Mastery Learning Test Model, which was developed for analyzing criterion referenced test data, is described. An estimate of the proportion of students in an instructional group which has achieved the referent objective is usable as a prior probability in interpreting individual responses. Considering instructional group performance enhances estimates of individual performance. Correlational data from a set of test items and a representative population of students are used to estimate the required item parameters. (Author)

FILMED FROM BEST AVAILABLE COPY

ED 076658

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
OFFICE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL OFFICE OF EDUCATION POSITION OR POLICY

3.11

Using Group Performance to Interpret Individual Responses to Criterion-Referenced Tests

Ronald Besel

Southwest Regional Laboratory for
Educational Research and Development
Los Alamitos, California 90720

Presented at the 1973 Annual Meeting of
the American Educational Research Association

February 1973, New Orleans

Using Group Performance to Interpret Individual Responses to Criterion-Referenced Tests

There is currently considerable controversy over what constitutes a criterion-referenced test. Typically the concept "criterion-referenced" is defined in relation to norm-referenced or standardized tests. For example... "norm-referenced measures compare the student's performance with the mean of a norm group whereas criterion-referenced measures compare his performance with a specified criterion score." (Livingston, 1972). On the basis of such definitions it has been claimed that interpretation of a student's performance on a criterion-referenced test should not be dependent upon the performance of his classmates or other norm groups.

> "the interpretation of a student's performance in a criterion-referenced situation is absolute and axiomatic, not dependent upon how other learners perform."
> (Airasian and Madaus, 1972)

> "(criterion-referenced)... measurements are absolute indices designed to indicate what the pupil has or has not learned from a given instructional segment. The measurements are absolute in that they are interpretable solely vis-a-vis a fixed performance standard or criterion and need not be interpreted relative to other measurements."
> (Block, 1971)

It is contended here that norm-group performance is useful and legitimate information for both the construction and application of criterion-referenced tests.

A criterion-referenced test is here defined as a set of items sampled from a domain which has been judged to be an adequate representation of an instructional objective. This definition does not limit criterion-referenced tests to narrowly defined behavioral objectives for which an item form (Osburn, 1968) specifies how to generate every item in the domain. It is desirable that the domain be described in operational terms; using this description another test developer should be able to generate an equivalent domain of test items. The assumptions or theory relating the domain of items to the referent objective should be explicitly stated.

Desirable procedures for selecting a sample of items from a domain depend upon the intended application of the test. One application of a criterion-referenced test is to estimate the proficiency of individual students relative to some achievement continuum. (Kriewall, 1972). This appears to be Glaser's (1963) original conception of the purpose of a criterion-referenced test. This application is based on the assumption that, "Underlying the concept of achievement measurement is the notion of a continuum of knowledge acquisition ranging from no proficiency at all to perfect performance." (Glaser, 1963). For applications where hand scoring of tests is used, a random or stratified random sampling of items from the domain permits the unweighted number-of-correct-responses to be interpreted as a degree of proficiency measure. If computer scoring is used, a sample of highly discriminating items will yield a better estimate of proficiency. Thus, the rejection of sampling based on item discrimination indices (norm-group performance) is based on the assumptions that a degree-of-proficiency measure is required and that the test must be hand-scored.

A frequent application of criterion-referenced tests is the making of categorical mastery, non-mastery decisions for students comprising an instructional group. Subsequent instruction for a student is contingent upon the category in which he is placed. Typically, test developers have computed a degree-of-proficiency index and then, on most frequently an arbitrary basis, selected a critical "passing" score. A problem that arises is that it is difficult, perhaps impossible, to define a meaningful degree-of-proficiency index for many types of legitimate instructional objectives. Ebel (1971) concludes that "criterion-referenced measurement may be practical in those few areas of achievement which focus on cultivation of a high degree of skill in the exercise of a limited number of abilities." Ebel's conclusion is based on the premise that a degree-of-proficiency scale "...anchored at the extremities--a score at the top of the scale indicating complete or perfect mastery of some defined abilities; one at the bottom indicating complete absence of those abilities." is required. Fortunately, such a measurement scale is not needed for the categorical decision application.

The Mastery Learning Test Model

The Mastery Learning Test Model has been designed to provide an appropriate algorithm for analyzing criterion-referenced test data for making the following instruction decision: "which students have achieved the referent objective." Two statistics are computed: the probability that a given student has achieved the objective and the proportion of an instructional group that have achieved the objective. The model assumes that each student in an instructional group can be treated as belonging to one of two groups--a group that has achieved the objective and one that has failed to achieve. The two-state assumption does not deny the possibility of partial achievement

of the objective. It does imply that categorization of students into two groups, masters and non-masters, is the desired type of decision and the basis for subsequent instruction.

The Mastery Learning Test Model and the true score theory upon which it is based are derived in an earlier paper (Besel, 1972). This model is related to a simpler mastery testing model suggested by Emrick (1971). Emrick's model assumes that measurement error can be accounted for by two test parameters: α -- the probability that a non-master will give a correct answer to an item; and β -- the probability that a master will given an incorrect answer to an item. His model implicitly assumes that all item difficulties and inter-item correlations are equal. This assumption can be avoided by increasing the number of test parameters--either by permitting item α parameters, or item β parameters, or both.

Parameter Estimation

Both the α and β item parameters can be estimated from the item response data collected from a representative sample of students. Two parameter estimation algorithms have been developed (Besel, 1973, a and b) for a Mastery Learning Model which has a single test--β parameter and item--α parameters. Least squares estimates of the parameters are computed using three classes of empirical data:

1. Item difficulties,
2. Inter-item covariances,
3. Score histograms.

The first algorithm (Besel, 1973 a) computes the least-squares estimates using an independent estimate of the proportion of student that have achieved the referent objective (GMP). The second algorithm requires no input estimate of GMP: it is estimated from the data in addition to the α parameters.

The stability of the parameter estimates was evaluated, for each algorithm, using test data from the end-of-unit Criterion Exercises of the SWRL Beginning Reading Program. Data from two consecutive years (1970-71 and 1971-72) were sampled from schools participating in the Quality Assurance Tryout. Each Criterion Exercise measures the achievement of four objectives: (1) Storybook Words, (2) Program Word Elements, (3) Word Attack (novel words), and (4) Letter Names. Five, three-option, multiple-choice items are used for each objective. Data from all ten units of the program were analyzed; the sample sizes shrank from 263 to 98 for the first year and from 418 to 173 for the second year.

The means and variances of the differences between the parameter estimates for the two years were examined (see Table 1). Computations were made for item α, average α ($\bar{\alpha}$), and test β. For the "Fixed GMP" algorithm two estimates of GMP were used. The first estimate was the proportion of students scoring 80% (4 right out of 5) or better for the outcome. The second estimate was the proportion with a perfect score. The item α differences are based on 50 items, average α and test β on 10 tests. The mean differences could be due partially to systematic differences in the student populations. Different school districts were represented in the two samples. The variances are more appropriate estimates of parameter stability.

For the second algorithm (GMP not fixed) the variances vary considerably across outcomes. The "fixed-GMP" algorithm achieved uniformly better stability with the perfect score criterion noticeably better than with the 80% criterion. The variances for both item α and average α decreased as the difficulty of the objective increased. Letter names is the easiest objective, word attack the most difficult. The variances of test β, on the other hand, increased as the difficulty of the objective increased. This trend was apparent in all three sets of calculations for both algorithms. This result is consistent with the notion that ideally one would like to estimate β from the responses of a group--all of which have achieved the objective. Likewise the item alphas could be "best" estimated from a group--none of which have achieved the objective. When a mixed group is used, β is estimated most accurately when a high proportion of the group has achieved the objective. Lowering the GMP of the norm group improves the accuracy of the α estimates at the expense of β accuracy.

Prior Probabilities

The Mastery Learning Model is a Bayesian statistical model. The response of a student to an item from the test is used to modify an existing probability estimate that the student has achieved the referent objective. A Bayesian model requires an initial prior probability estimate. One estimate which results in better probability-of-mastery estimates than an "ignorance" (prior probability equal .5) assumption is the estimated proportion of students in mastery (GMP) for the appropriate instructional group. If the test parameters have been previously estimated for a representative norm group, Equation 1 (Besel, 1972), can be used to estimate GMP.

$$GMP = \frac{U/K - \bar{\alpha}}{1-\bar{\alpha}-\bar{\beta}} \qquad (1)$$

GMP = proportion of students in mastery state

$\bar{\alpha}$ = average of item α parameters

$\bar{\beta}$ = average of item β parameters or test β parameter

U/K = mean percentage score

While the use of group-estimated priors is somewhat controversial for selection decisions across instructional groups (Novick, 1970), it promises to enhance instructional decisions within an instructional group.

Summary

The usage of an independent estimate of the proportion of students in a norm group which have achieved an objective resulted in significantly improved stability of Mastery Learning parameters. This should result in increased validity of the Mastery Learning Test Model for making categorical mastery--non-mastery decisions. This Test Model can be used to make mastery decisions on the basis of very short tests. Using the proportion-in-mastery estimate for an instructional group as a prior-probability results in improved estimates of the probability that an individual student has achieved the objective. Norm group data can also be used to select the best set of items from a domain for the mastery decision application.

Table 1. Stability of Mastery Learning Parameters
(Mean Difference/Variances of Difference)

Outcome	Parameter	Minimum Sum of Squares Solution	80% Criterion Solution	100% Criterion Solution
1	Item α	-.081 / .0276	-.026 / .0191	-.013 / .0076
	$\bar{\alpha}$	-.081 / .0122	-.026 / .0031	-.013 / .0008
	β	.018 / .0006	-.002 / .0002	-.004 / .0001
2	Item α	-.059 / .0126	-.042 / .0170	-.041 / .0072
	$\bar{\alpha}$	-.059 / .0033	-.042 / .0015	-.041 / .0004
	β	-.003 / .0004	-.007 / .0005	-.006 / .0001
3	Item α	-.037 / .0083	-.032 / .0096	-.020 / .0043
	$\bar{\alpha}$	-.037 / .0011	-.032 / .0017	-.020 / .0007
	β	-.000 / .0006	-.001 / .0006	-.003 / .0001
4	Item α	..052 / .0956	-.026 / .0354	-.036 / .0080
	$\bar{\alpha}$	.052 / .0418	-.026 / .0101	-.036 / .0010
	β	-.004 / .0002	-.006 / .0002	-.004 / .0000

References

Airasian, P.W. and Madaus, G.F. "Criterion-Referenced Testing in the Classroom." Measurement in Education, 3 (May, 1972).

Besel, R.R. "A Mastery Learning Test Model." Presented at the 1972 Annual Meeting of the American Educational Research Association, (April, 1972).

Besel, R.R. "Program for Computing Least Squares Estimates of Item Parameters for the Mastery Learning Test Model: Fixed GMP Version," SWRL 1973(a).

Besel, R.R. "Program for Computing Least Squares Estimates of Item Parameters for the Mastery Learning Test Model: Variable GMP Version," SWRL 1973(b).

Block, J.W. "Criterion-Referenced Measurements: Potential." School Review, 79 (1971), 289-298.

Ebel, R. L. "Criterion-Referenced Measurements: Limitations." School Review, 79 (1971), 282-288.

Emrick, J.A. "An Evaluation Model for Mastery Testing." Journal of Educational Measurement, 8 (Winter, 1971), 321-326.

Glaser, R. "Instructional Technology and the Measurement of Learning Outcomes: Some Questions." American Psychologist, 18 (1963), 519-521.

Kriewall, T.E. "Aspects and Applications of Criterion-Referenced Tests." Presented at the 1972 Annual Meeting of the American Educational Research Association, (April, 1972).

Livingston, S.A. "Criterion-Referenced Applications of Classical Test Theory." Journal of Educational Measurement, 9 (Spring, 1972), 13-26.

Novick, M.R. "Bayesian Considerations in Educational Information Systems" in Proceedings of the 1970 Invitational Conference on Testing Problems. Educational Testing Service, (October, 1970), 70-88.

Osburn, H. G. "Item Sampling for Achievement Testing." Educational and Psychological Measurement, 28 (1968), 95-104.